LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models

Year: 2023 | *Citations: 207* | *Authors: Long Lian, Boyi Li, Adam Yala*

Abstract

Recent advancements in text-to-image diffusion models have yielded impressive results in generating realistic and diverse images. However, these models still struggle with complex prompts, such as those that involve numeracy and spatial reasoning. This work proposes to enhance prompt understanding capabilities in diffusion models. Our method leverages a pretrained large language model (LLM) for grounded generation in a novel two-stage process. In the first stage, the LLM generates a scene layout that comprises captioned bounding boxes from a given prompt describing the desired image. In the second stage, a novel controller guides an off-the-shelf diffusion model for layout-grounded image generation. Both stages utilize existing pretrained models without additional model parameter optimization. Our method significantly outperforms the base diffusion model and several strong baselines in accurately generating images according to prompts that require various capabilities, doubling the generation accuracy across four tasks on average. Furthermore, our method enables instruction-based multi-round scene specification and can handle prompts in languages not supported by the underlying diffusion model. We anticipate that our method will unleash users' creativity by accurately following more complex prompts. Our code, demo, and benchmark are available at: https://llm-grounded-diffusion.github.io